MACHINE LEARNING

Field of study that gives computers the ability to learn without being explicitly programmed.

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

The key difference between traditional programming and machine learning is while you set data and program following an output, in machine learning, you set data and output to receive a program making predictions.

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

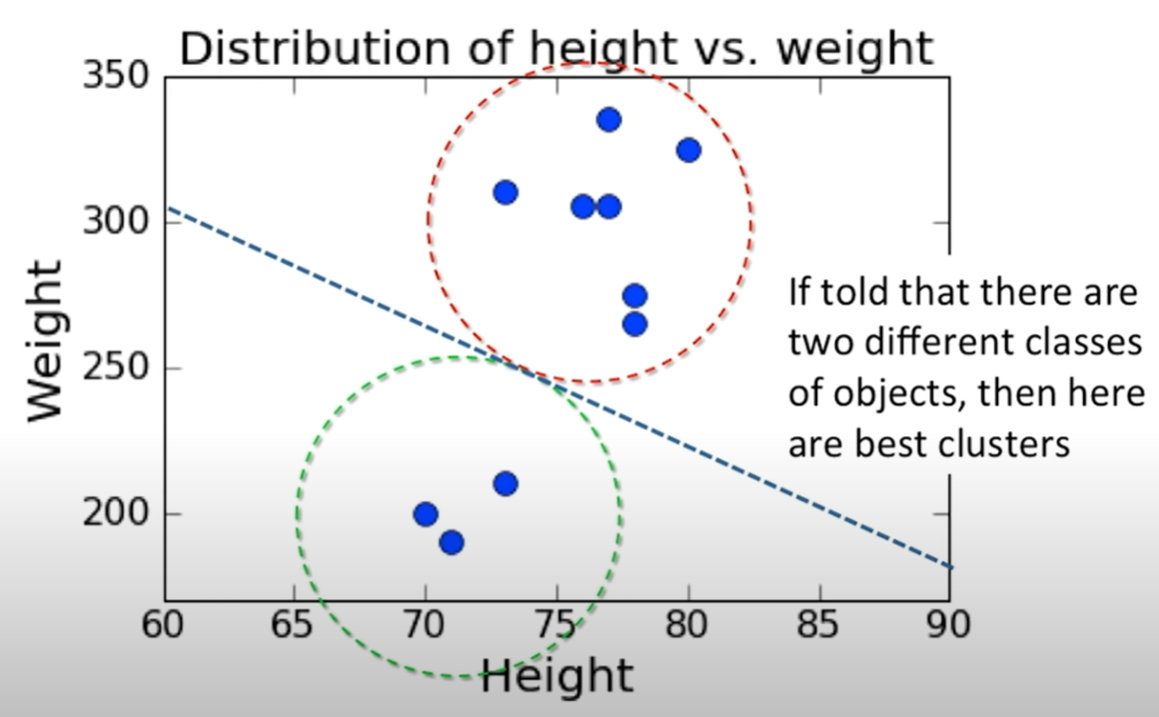
**Example of basic paradigm:**

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Clustering:**

Grouping set of objects having similar value in a 2d dimensional graph. Objects are often represented as dots by representing two feature values (on x and y-axis). Objects having similar positions are covered within a circle and the process is called clustering.



**Overfitting:**

It is a non-desirable outcome happening when a machine learning algorithm learns the training data too well. This causes it captures noise or random fluctuations in the data rather than the underlying patterns. Ultimately, the model performs well on training data but poorly on new, unseen data because of memorization of training data. Can be understood by having high accuracy on training data but low on test data.

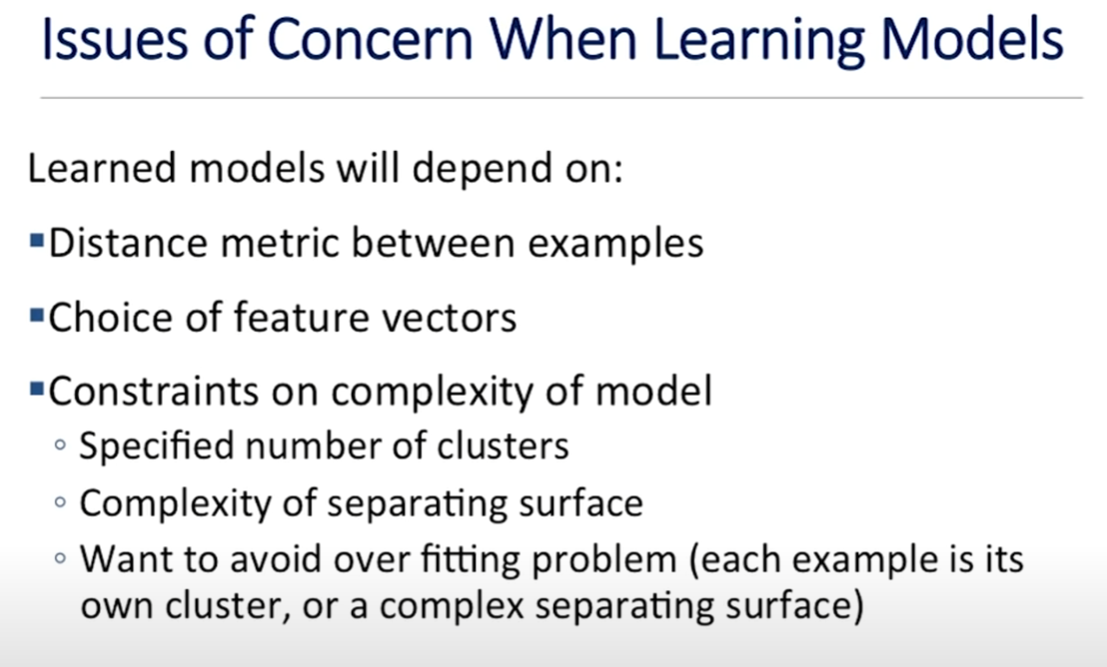
Inserting too many features can cause overfitting.

**False Positives and False Negatives:**

In false negative, the model incorrectly predicts positive outcome when the true outcome is negative, it is vice versa for false positive. Consider positive outcome as an email classified as spam and negative outcome an email that is not classified as spam.

**Signal to Noise Ratio:**

Maximize the categories having the most information and remove which doesn’t.



metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

As a more complex surfaces, sequence of lines to separate data can be used but it must be cautious to not overfit the data.

**Positive predictive value = true positive / (true positive + false positive)**

**Sensitivity = true positive / (true positive + false negative) 🡪** percentage correctly found

**Specificity = true negative / (true negative + false positive) 🡪** percentage correctly rejected

**Kth-nearest Neighbors Algorithm:**

It is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. To implement the algorithm, distance between the selected point and the test data is calculated, then k number of closest elements are taken. At the end, selected point is assessed as the label of the chosen data that contains same label the most. Note that the Euclidean distance (square hood of subtracted squared equation) is calculated.

metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

While selecting value of k, it can be looked at the square root of n where n is the total number of data points. Make sure that k is odd to avoid confusion between two classes of data.

KNN algorithm is efficient to use when the dataset is small, data is noise free (no random data in specific field) and it is labeled.